HOUSE PRICE PREDICTION

by Arpan KANANI

Submission date: 27-Apr-2021 05:34PM (UTC+0530)
Submission ID: 1433559952
File name: HOUSE_PRICE_PREDICTION_USING_REGRESSION_TECHNIQUES.docx (694.34K)
Word count: 1657
Character count: 8434

HOUSE PRICE PREDICTION USING REGRESSION TECHNIQUES

ARPAN KANANI

Pursuing B. Tech(IT), CSPIT

Charotar University of Science and Technology

Anand, India

ABSTRACT

The main part of any business problem is considered to be building models and train and tweak them to create beautiful solutions. In fact, most of the data scientist spend their time on turning business problem into data problems and after that collection of data and understanding, cleaning and formatting of data which is as an input to machine learning algorithms afterthough. Despite, it is interesting and essential to built models and obtain desired solutions. Here we will be using Kaggle dataset - Bangalore price prediction, to understand various regression techniques such as linear regression, lasso regression, support vector machine(SVM), random forest regressor, and finally keras regression which is a neural network technique. We will also be discussing the cost function of various algorithms and later evaluation techniques such as mean square error, r2_score to evaluate model accuracy.

INDEX WORDS: Linear regression, Lasso, SVM, MSE, RMSE, Keras Regression.

INTRODUCTION

Each and every individual has the need of roof over their heads. this roof could be varied forms such as a home, a mention, a apartment and many more. The real problem arises when someone wishes to buy a house of their choice but get confused and in the end compromise their choices due to certain reasons. One of them and the peculiar one is cost of the house, which are getting expensive day-by-day. And why so ?

We have observed that many a times, people don't intentionally buy the house for their needs, rather than for growing their business and investment purpose to gain more profit. Moreover, actual price of the house is not-known and the deal is fixed. However, it is always better to get prior knowledge about the cost of house and how much price is worthful to particular property. Additionally, it is even more significant to find out the cause for hype of price.

Here, in this work we would be searching for the answer on various factors which are adequate. let's begin with some questionaries and try to grasp and be familiar with problems with the house pricing.

- Lack of knowledge with the seller regarding how to hype the house price really to gain profit.(e.g. while renovation the house building a pool could be more valuable and profitable than repairing the bathroom)
- The seller might have very few knowledge about the actual or adequate price of house.
- The buyer lacks knowledge of how much the property cost.
- Commercial problem: auction services may be interested in tools to support sellers and buyers (to highlight the sections in the offers that most affect the price).[1]

DESIGN APPROACH

Here, we have tried to implement varied form of regression algorithm to get the best accuracy and low mean square error:

A. Linear Regression:

Linear regression is process of predicting continuous value. More generally, a linear model makes a prediction by simply computing sum of input features, plus a constant called bias term (intercept term).[2]

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

In this equation:

- $\hat{y}$ is the predicted value.
- n is the number of features.
- x_i is the i^{th} feature value.
- θ_j is the j^{th} model parameter (including the bias term θ and the feature weights θ_1, θ_2 , .., θ_n).

The MSE of a Linear Regression hypothesis h on a training set X is calculated using below equation:

$$\text{MSE}(\mathbf{X}, h_{\boldsymbol{\theta}}) = \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{\theta}^{\intercal}\mathbf{x}^{(i)} - y^{(i)}\right)^2$$

B. Lasso Regression

Least Absolute Shrinkage and Selection Operation Regression (usually called Lasso regression) is another regularized version of linear regression: just like Ridge Regression, it adds a regularization term to the cost function, but it uses the L1 norm of the weight vector instead of half the square of the L2 norm.[2]

$$J(\boldsymbol{\theta}) = \text{MSE}(\boldsymbol{\theta}) + \alpha\sum_{i=1}^{n}|\theta_i|$$

An important characteristic of Lasso Regression is that it tends to eliminate the weights of the least important features. In other words, Lasso Regression automatically performs feature selection and outputs a sparse model (i.e.,with few nonzero feature weights).

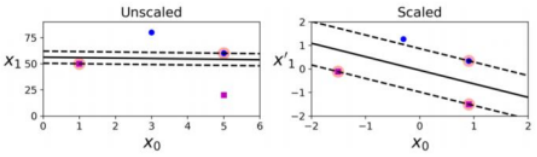
There are two main differences with Lasso. First, the gradients get smaller as the parameters approach the global optimum, so Gradient Descent naturally slows down, which helps convergence (as there is no bouncing around). Second, the optimal parameters (represented by the red square) get closer and closer to the origin when you increase α, but they never get eliminated entirely.

$$g(\boldsymbol{\theta}, J) = \nabla_{\boldsymbol{\theta}}\,\text{MSE}(\boldsymbol{\theta}) + \alpha\begin{pmatrix}\text{sign}(\theta_1)\\ \text{sign}(\theta_2)\\ \vdots\\ \text{sign}(\theta_n)\end{pmatrix} \quad \text{where } \text{sign}(\theta_i) = \begin{cases}-1 & \text{if } \theta_i < 0\\ 0 & \text{if } \theta_i = 0\\ +1 & \text{if } \theta_i > 0\end{cases}$$

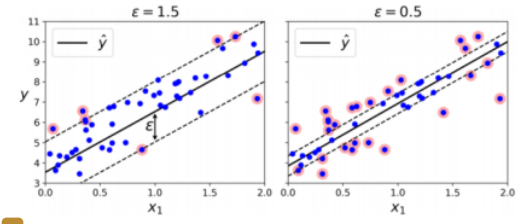
C. Support Vector Machines

A Support Vector Machine (SVM) is a powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification, regression, and even outlier detectionSVM is supervised algorithm that classifies cases by finding a separator.[2]

1. Mapping data to a high dimensional feature space.
2. Finding a separator.



To use SVMs for regression instead of classification, the trick is to reverse the objective: instead of trying to fit the largest possible street between two classes while limiting margin violations, SVM Regression tries to fit as many instances as possible on the street while limiting margin violations (i.e., instances off the street). The width of the street is controlled by a hyperparameter, ϵ.



We can use Scikit-Learn's SVR to perform SVM Regression.

Neural Networks

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.[3]

Since, we had 242 features (after on one hot encoding), so therefore we have inserted 252 neuron as a start, 4 hidden layers and a output layer to predict the price of the house. Also, Adam optimisation function and activation function "relu" and loss is measured in the form of mean square error (MSE).

```
# having 245 neuron is based on the number of available features
model = Sequential()
model.add(Dense(245,activation='relu'))
model.add(Dense(500,activation='relu'))
model.add(Dense(500,activation='relu'))
model.add(Dense(500,activation='relu'))
model.add(Dense(1))
model.compile(optimizer='Adam',loss='mse')
```

Further analysis of the model and calculating the accuracy of the model with variance score, we got 85.40% .

```
MAE: 17.062710305594063
MSE: 987.1336016752132
RMSE: 31.4186823669487
VarScore: 0.8540924744776565
```

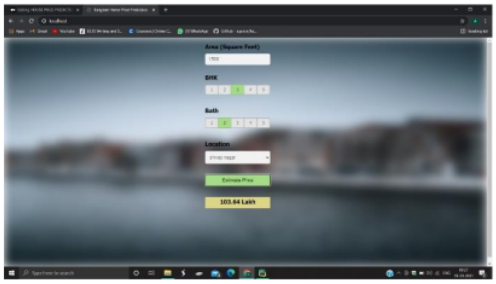
After getting analysis of the dataset we have derived the model with 85% accuracy so our next step would be to import model and its features in the form of pickle and json file respectively. This insight would be useful in further creating a interactive website. Moreover, our main goal is to host the website on amazon ec2 instace with the help of nginx server.

To freely interact our website with the model and derive our features (locations) we have created a flask server which would predict the prices of house in different location based on the attributes that are provided by the user as an input.

```
@app.route('/get_location_names', methods=['GET'])
def get_location_names():
    response = jsonify({
        'locations': util.get_location_names()
    })
    response.headers.add('Access-Control-Allow-Origin', '*')
    return response
```

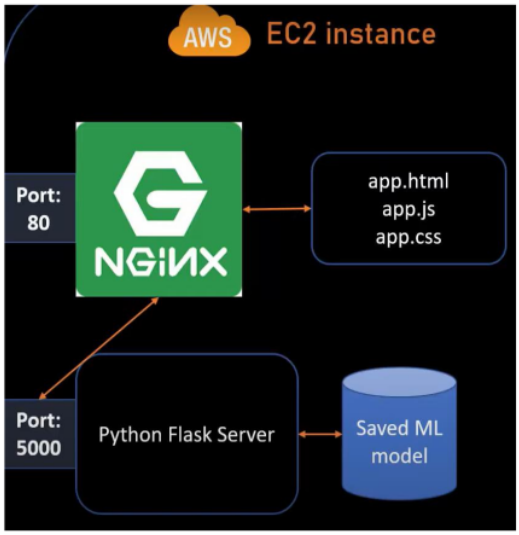
```
@app.route('/predict_home_price', methods=['GET','POST'])
def predict_home_price():
    total_sqft = float(request.form['total_sqft'])
    location = request.form['location']
    bhk = int(request.form['bhk'])
    bath = int(request.form['bath'])

    response = jsonify({
        'estimated_price': util.get_estimate_price(location,total_sqft,bhk,bath)

    })
    response.headers.add('Access-Control-Allow-Origin', '*')
    return response
```

Here is the output of our newly created website which has simple architecture as our end goal is to predict the price of house on the website which we host on aws ec2 instance.



Architecture of EC2 instance

Nginx is the light web server which can serve http requests. Here, we will be using nginx as web server. When we load the url in the web browser it will direct to our ec2 instance and from there our nginx server will load our webpages and it related files. Moreover, when we click on button to predict the price from the browser from the javascript file, again we will call the nginx server and further use a reverse proxy setup to route our api requests to the python flask server which we have created and is running on the same ec2 instance which is using our derived model to cater prediction requests.



Proposed steps of hosting a website on EC2 instance with the help of Nginx server:

1. Launch EC2 instance on Amazon web service, however choose Ubuntu machine with free tier.
2. Other steps of the instance would be default except for security group, we have to create our own security group and manually specify two rules (Http and Https).
3. Review and launch the instance. After that we need to download the key which is in .pem format.
4. Now connect to your instance using a command like this,

```
ssh -i "C:\Users\Viral\.ssh\Banglore.pem" ubuntu@ec2-3-133-88-210.us-east-2.compute.amazonaws.com
```

5. nginx setup

i. Install nginx on EC2 instance using these commands,

```
sudo apt-get update
sudo apt-get install nginx
```

ii. Above will install nginx as well as run it. Check status of nginx using

```
sudo service nginx status
```

iii. Here are the commands to start/stop/restart nginx

```
sudo service nginx start
sudo service nginx stop
sudo service nginx restart
```

iv. Now when you load cloud url in browser you will see a message saying "welcome to nginx" This means your nginx is setup and running.

6. Now you need to copy all your code to EC2 instance. You can do this either using git or copy files using winscp. We will use winscp. You can download winscp from here: https://winscp.net/eng/download.php.
7. Once you connect to EC2 instance from winscp (instruction in a youtube video), you can now copy all code files into /home/ubuntu/ folder. The full path of your root folder is now: /home/ubuntu/Specified folder.
8. After copying code on EC2 server now we can point nginx to load our property website by default. For below steps,
   i. Create this file /etc/nginx/sites available/bhp.conf. The file content looks like this,

```
server {
    listen 80;
        server_name bhp;
        root /home/ubuntu/BangloreHomePrices/client;
        index app.html;
        location /api/ {
             rewrite ^/api(.*) $1 break;
             proxy_pass http://127.0.0.1:5000;
        }
}
```

   ii. Create symlink for this file in /etc/nginx/sites-enabled by running this command,

```
sudo ln -v -s /etc/nginx/sites-available/bhp.conf
```

   iii. Remove symlink for default file in /etc/nginx/sites-enabled directory,

```
sudo unlink default
```

   iv. Restart nginx,

```
sudo service nginx restart
```

9. Now install python packages and start flask server
10. Now just load your cloud url in browser (for me it was ec2-13-59-50-29.us-east-2.compute.amazonaws.com) and this will be fully functional website running in production cloud environment.

Future work

In our proposed model the RMSE score achieved is comparatively low, but improvements can be made. In real world data, we can use such a model but with the better dataset and good amount of features to come up with a better result. Furthermore, updation is necessary to expand the dataset and relate with the current timing.

References

[1]. Piotr Grązka (SGH Warsaw School of Economics), Anna Kozak (Warsaw University of Technology), Paweł Wicherek (Warsaw University of Technology).

[2]. Aurelien Geron (Hands-on lab Machine learning with Scikit-Learn, Keras and Tensor-Flow).

[3] Masha Mir (towards datascience: House Prices Prediction Using Deep Learning).

[4] Romain Strock(Static website hosting with nginx on EC2,blog).

HOUSE PRICE PREDICTION

ORIGINALITY REPORT

25%	24%	15%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.oreilly.com Internet Source	9%
2	index-of.es Internet Source	7%
3	www.slideshare.net Internet Source	4%
4	www.coursehero.com Internet Source	3%
5	www.lpsm.paris Internet Source	1%
6	jingwen-z.github.io Internet Source	1%
7	Joaquim Carreras, Yara Yukie Kikuti, Masashi Miyaoka, Shinichiro Hiraiwa et al. "A Combination of Multilayer Perceptron, Radial Basis Function Artificial Neural Networks and Machine Learning Image Segmentation for the Dimension Reduction and the Prognosis Assessment of Diffuse Large B-Cell Lymphoma", AI, 2021	<1%

Publication

8	builtin.com Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		